Министерство образования и науки Российской Федерации
Московский физико-технический институт
(государственный университет)
Факультет управления и прикладной математики
Вычислительный центр им. А. А. Дородницына РАН
Кафедра «Интеллектуальные системы»

Гончаров Алексей Владимирович

Методы дискретного и непрерывного выравнивания временных рядов

010900 - Прикладные математика и физика

Выпускная квалификационная работа бакалавра

Научный руководитель:
д.ф.-м.н. Стрижов Вадим Викторович

Москва
2016

Содержание

Аннотация

В работе рассматривается задача метрического анализа и классификации временных рядов. Метрические методы используют матрицу попарных расстояний, строящуюся при помощи фиксированной функции расстояния. Вычислительная сложность алгоритмов, использующих такую матрицу, по меньшей мере квадратична относительно числа временных рядов. Проблема снижения вычислительной сложности решается путем предварительного выделения эталонных объектов, центроидов классов, и последующего их использования для описания классов. В качестве базовой модели классификации выбрана модель, использующая динамическое выравнивание временных рядов для построения центроида. Предлагается ввести функцию весов центроида, влияющую на вычисление расстояния между объектами. Также существуют задачи, связанные с обработкой временных рядов, частота измерений которых различна. При решении подобных задач удобно считать объекты непрерывными. Но стандартные методы метрического анализа, такие как функция расстояния DTW, определены лишь для случая дискретных временных рядов. Область применения техники динамического выравнивания временных рядов расширяется на непрерывный случай. При этом точный выравнивающий путь найти невозможно из-за невозможности перебора путей. Ищется и используется его аппроксимация. Такая аппроксимация должна быть устойчива как к небольшим отклонениям от своей траектории, так и к флуктуациям значений временного ряда. Данный подход не накладывает ограничений на вид аппроксимации временного ряда, а также на вид аппроксимации пути между временными рядами. Свойства построенных моделей исследуются и сравниваются со свойствами модели, выбранной в качестве базовой.

Ключевые слова: взвешенное динамическое выравнивание; центроид; функция расстояния; непрерывные временные ряды; выравнивание непрерывных объектов.

1 Введение

Рассматривается задача анализа и классификации временных рядов. Существуют различные способы ее решения: построение признакового пространства, использование нейронных сетей, аппроксимация параметрическими функциями. Так, в [1] исследованы методы построения признакового описания временных рядов, в частности метод экспертного построения признаков и метод построения признакового описания на основе гипотезы порождения данных. Результаты [1] показывают, что построенное признаковое пространство адекватно описывает зависимую переменную. В [2] для решения задачи классификации использованы нейронные сети с небольшим числом связей между нейронами, обладающие свойством устойчивости к возмущениям данных. В [3] для классификации предложен алгоритм разбиения исходных временных рядов на периоды и их очистки от шумов. Предложены модификации алгоритма k ближайших соседей и нейронной сети для решения поставленной задачи. В вычислительном эксперименте, проведенном на реальных данных, оценена эффективность, а также проведено сравнение данных алгоритмов между собой. При этом показан высокий процент правильной классификации.

Построение матрицы попарных расстояний между всеми объектами в задаче метрической классификации является вычислительно трудоемкой задачей. Для снижения размерности задачи и вычислительных затрат предлагается решать задачу с предварительным выделением эталонных объектов, или же центроидов классов, и последующим их использованием для описания множества временных рядов.

Метрические методы используют различные функции расстояния для построения матрицы попарных расстояний: евклидово расстояние [4], метод динамического выравнивания временных рядов [5, 6], метод, основанный на нахождении наибольшей общей последовательности [7], Edit Distance with Real Penalty [8], Edit Distance on Real sequence [9], DISSIM [10], Sequence Weighted Alignment model [11], Spatial Assembling Distance [12] и другие. В качестве базового метода для построения функции расстояния в настоящей работе предлагается использовать динамическое выравнивание временных рядов (Dynamic Time Warping) [13]. Как показано в [14], этот метод находит наилучшее соответствие между двумя временными рядами, если они нелинейно деформированы друг относительно друга — растянуты, сжаты или смещены вдоль оси времени.

Базовой моделью классификации и анализа временных рядов в текущей работе принята модель, описанная в [15]. Там в качестве центроида выбирается объект выборки, являющийся ближайшим ко всем остальным объектам. Применяться же будет метод точного его вычисления. Это метод DBA, решающий задачу оптимизации для нахождения центроида. Алгоритм применения и доказательство корректности приведены в [16].

Последующая работа разделена на два раздела. Первый посвящен построению функции расстояния между дискретными временными рядами, а второй — между непрерывными. Опишем каждый из них более конкретно.

В следующем разделе вводится понятие вектора весов и матрицы весов центроида и описываются методы vwDTW и mwDTW (векторно- и матрично-взвешенный DTW) вычисления функции расстояния, основывающиеся на следующем предположении о форме временных рядов: в одном и том же классе находятся временные ряды, имеющие схожую форму с точностью до линейной или нелинейной деформации, локальных или глобальных сдвигов по оси времени. Предполагается, что в центроиде присутствуют характерные для всего класса участки, которым соответствуют большие веса вектора весов этого центроида. А функция расстояния, основанная на vwDTW и использующая вектор весов центроида, точнее объединит объекты одного класса и разделит объекты разных классов, чем основанная на DTW. Поэтому в предлагаемой модели используется метод vwDTW как для вычисления центроида по методу DBA, так и для построения матрицы попарных расстояний. Также исследуются свойства, вид матрицы весов центроида и эффективность применения расстояния, вычисленного с ее помощью — mwDTW, в прикладных задачах.

Для дальнейшей классификации рядов по полученной матрице расстояний сниженной размерности применяется метод k ближайших соседей, как и в базовой модели. Процедура классификации выполняется в три шага. Первый — отбор эталонных объектов каждого класса. Второй — построение матрицы попарных расстояний сниженной размерности между временными рядами и эталонными объектами каждого класса. Третий — классификация временных рядов методом k ближайших соседей с помощью матрицы попарных расстояний.

Для проверки работоспособности такой модели проведен вычислительный эксперимент на реальных и синтетических данных. Экспери-

мент включает в себя анализ и классификацию данных при помощи построенной модели. Полученные результаты сравниваются с результатами применения базовой модели к тем же исходным данным.

Расстояние DTW определено лишь между дискретными временными рядами. При замене этих объектов непрерывными аналогами исчезает проблема различий в частоте измерений, но стоит проблема применимости функции расстояния. В этой работе вводится понятие функции расстояния DTW между непрерывными временными рядами, пути между ними, стоимости пути и выравнивающий пути. В крайнем разделе границы применимости метода расширены на непрерывный случай.

В дискретном случае поиск выравнивающего пути осуществляется при помощи динамического программирования [13]. В непрерывном же случае воспользоваться перебором невозможно, так как множество путей несчетно. Проблема поиска выравнивающего пути решена путем аппроксимации реального пути параметрической функцией. В качестве класса параметрических функций может быть выбран любой класс, например различные сплайны [18]. Поиск пути таким образом сводится к поиску оптимальных параметров, задающих его приближение.

Для проведения вычислительного эксперимента использованы временные ряды акселерометра мобильного телефона. Изначально они представляют собой дискретный временной ряд ускорения телефона в трех координатах. По этим данным создаются непрерывный объекты. Это также делается через аппроксимацию исходного временного ряда параметрическими функциями. В работе используется интерполяция кубическими сплайнами.

Универсальность данного подхода заключается в возможности применять любой способ аппроксимации временного ряда, а также произвольный способо аппроксимации пути наименьшей стоимости при возможности наложения соответствующих ограничений.

Таким образом, в разделе создается функция расстояния между непрерывными временными рядами. Исследуются свойства выравнивающего пути между ними и его стоимости.

2 Взвешенное динамическое выравнивание

2.1 Постановка задачи

Задана выборка $\mathfrak{D} = \{(\mathbf{s}_i, y_i)\}_{i=1}^m$, состоящая из пар объект — ответ. Объектами служат временные ряды $\mathbf{s}_i \in \mathbb{R}^n$, а ответами являются метки класса — $y_i \in Y = \{1, ..., E\}$, где $E \ll m$. Выборка разбита на обучающую $\mathfrak{D}_l$ и контрольную $\mathfrak{D}_t$.

Определение 1. *Модель классификации f — параметрическая функция объектов выборки, приближающая целевую зависимость y_i. В данной работе параметрами модели примем множество центроидов $\boldsymbol{C} = \{\boldsymbol{c}_e\}_{e=1}^E$ и множество векторов весов центроидов $\hat{\boldsymbol{W}} = \{\boldsymbol{w}_e\}_{e=1}^E$ или же матриц весов центроидов $\hat{\boldsymbol{W}} = \{\boldsymbol{W}_e\}_{e=1}^E$.*

Определение 2. *Функцией ошибки S модели f для задачи классификации будем считать*

$$S(f, \mathfrak{D}_t) = \frac{1}{|\mathfrak{D}_t|} \sum_{i=1}^{|\mathfrak{D}_t|} [f(\boldsymbol{s}_i) \neq y_i].$$

Требуется построить модель классификации $f : \mathbb{R}^n \to Y$, минимизирующую функцию ошибки S на контрольной выборке:

$$f_{\mathbf{C},\hat{\mathbf{W}}} = \underset{\mathbf{C},\hat{\mathbf{W}}}{\operatorname{argmin}} \big(S(f, \mathfrak{D}_t)\big). \tag{1}$$

2.2 Вычисление значения функции расстояния

2.2.1 Общие понятия

В данной работе в качестве метрического расстояния между объектами предлагается использовать стоимость *взвешенного пути наименьшей стоимости* между этими объектами.

Даны два временных ряда: $\mathbf{s}_1$ и $\mathbf{s}_2$. Будем считать, что $\mathbf{s}_1, \mathbf{s}_2 \in \mathbb{R}^n$. Пусть $\boldsymbol{\Omega}^{n \times n}$ — это матрица, такая что ее элемент Ω_{ij} равен квадрату разности между i-м и j-м элементами последовательностей $\mathbf{s}_1$ и $\mathbf{s}_2$:

$$\Omega_{ij} = (s_{1i} - s_{2j})^2.$$

Определение 3. *Путем $\boldsymbol{\pi}$ между последовательностями $\boldsymbol{s}_1$ и $\boldsymbol{s}_2$ назовем упорядоченное множество пар индексов элементов матрицы*

$\boldsymbol{\Omega}$:

$$\boldsymbol{\pi} = \{\pi_r\} = \{(i_r, j_r)\}, \quad r = 1, \ldots, R, \quad i, j \in \{1, ..., n\},$$

где R — длина пути, зависящая от выбора пути. Он должен удовлетворять следующим условиям.

Граничные условия. *$\pi_1 = (1, 1)$ и $\pi_R = (n, n)$, т. е. начало и конец $\boldsymbol{\pi}$ находятся на диагонали в противоположных углах $\boldsymbol{\Omega}$.*

Непрерывность. *Пусть $\pi_r = (p_1, p_2)$ и $\pi_{r-1} = (q_1, q_2), \quad r = 2, ..., R$. Тогда*

$$p_1 - q_1 \leq 1, \quad p_2 - q_2 \leq 1.$$

Это ограничение нужно, чтобы в шаге пути $\boldsymbol{\pi}$ участвовали только соседние элементы матрицы (включая соседние по диагонали).

Монотонность. *Пусть $\pi_r = (p_1, p_2)$ и $\pi_{r-1} = (q_1, q_2), \quad r = 2, ..., R$. Тогда выполняется хотя бы одно из условий*

$$p_1 - q_1 \geq 1, \quad p_2 - q_2 \geq 1.$$

Это ограничение обусловлено природой рассматриваемых последовательностей и предназначено для монотонности функции выравнивания времени.

Физические ограничения. *Как уже говорилось во введении, предполагается, что временные ряды одного класса имеют схожую форму и являются линейно или нелинейно деформированными или же смещены друг относительно друга. При этом считается, что подобного рода деформации и смещения являются малыми, локальными. В этом предположении выравнивающий путь в матрице слабо отклоняется от диагонали, то есть*

$$\text{для каждого} \quad \{i_r, j_r\} \in \boldsymbol{\pi} \quad i_r - k \leq j_r \leq i_r + k,$$

где k определяется типом задачи и ее физическими ограничениями.

2.2.2 Векторно-взвешенный путь наименьшей стоимости

Дадим определение векторно-взвешенного пути наименьшей стоимости, vwDTW. Дан вектор весов $\mathbf{w} \in \mathbb{R}^n$.

Определение 4. *Стоимостью $Cost(\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{w}, \boldsymbol{\pi})$ векторно-взвешенного пути $\boldsymbol{\pi}$ между последовательностями $\boldsymbol{s}_1$ и $\boldsymbol{s}_2$ с весом $\boldsymbol{w}$*

назовем

$$Cost(\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{w}, \boldsymbol{\pi}) = \sum_{(i,j)\in\boldsymbol{\pi}} w_j \Omega_{ij}. \tag{2}$$

Определение 5. *Векторно-взвешенным путем наименьшей стоимости (векторно-взвешенным выравнивающим путем)* $\hat{\boldsymbol{\pi}}$ *между последовательностями* $\boldsymbol{s}_1$ *и* $\boldsymbol{s}_2$ *назовем взвешенный путь, имеющий наименьшую стоимость среди всех возможных векторно-взвешенных путей между последовательностями* $\boldsymbol{s}_1$ *и* $\boldsymbol{s}_2$*:*

$$\hat{\boldsymbol{\pi}} = \underset{\boldsymbol{\pi}}{argmin}\, Cost(\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{w}, \boldsymbol{\pi}). \tag{3}$$

Обозначим стоимость векторно-взвешенного выравнивающего пути между последовательностями $\mathbf{s}_1$ и $\mathbf{s}_2$ через $\rho(\mathbf{s}_1, \mathbf{s}_2, \mathbf{w}) = \mathrm{Cost}(\mathbf{s}_1, \mathbf{s}_2, \mathbf{w}, \hat{\boldsymbol{\pi}})$.

Для вычисления стоимости такого пути в данной работе используется модифицированный метод DTW — vwDTW (vector-weighted DTW). Согласно этому методу необходимо построить новую матрицу $\boldsymbol{\Gamma}$, элементы которой определяются следующим образом:

$$\Gamma_{1j} = w_j \Omega_{1j}, \quad \Gamma_{i1} = w_1 \Omega_{i1}, \quad i, j = 1, \ldots, n,$$

$$\Gamma_{ij} = w_j \Omega_{ij} + \min(\Gamma_{i,j-1}, \Gamma_{i-1,j}, \Gamma_{i-1,j-1}), \quad i, j = 2, \ldots, n.$$

Элемент Γ_{ij} матрицы $\boldsymbol{\Gamma}$ равен стоимости векторно-взвешенного выравнивающего пути между последовательностями $\{s_{1a}\}_{a=1}^{i}$ и $\{s_{2a}\}_{a=1}^{j}$.

В качестве значения функции расстояния между двумя объектами выберем стоимость векторно-взвешенного выравнивающего пути между ними (3):

$$\rho(\mathbf{s}_1, \mathbf{s}_2, \mathbf{w}) = \Gamma_{nn}. \tag{4}$$

Заметим, что при единичном векторе весов vwDTW эквивалентен обычному DTW, описание которого приведено в [15].

2.2.3 Матрично-взвешенный путь наименьшей стоимости

Дадим определение матрично-взвешенного пути наименьшей стоимости, mwDTW. Дана матрица весов $\mathbf{W} \in \mathbb{R}^{n\times n}$.

Определение 6. *Стоимостью* $Cost(\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{W}, \boldsymbol{\pi})$ *матрично-взвешенного пути* $\boldsymbol{\pi}$ *между последовательностями* $\boldsymbol{s}_1$ *и* $\boldsymbol{s}_2$ *с весом* $\boldsymbol{W}$ *назовем*

$$Cost(\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{W}, \boldsymbol{\pi}) = \sum_{(i,j)\in\boldsymbol{\pi}} W_{ij}\Omega_{ij}. \tag{5}$$

Определение 7. *Матрично-взвешенным путем наименьшей стоимости (матрично-взвешенным выравнивающим путем)* $\hat{\boldsymbol{\pi}}$ *между последовательностями* $\boldsymbol{s}_1$ *и* $\boldsymbol{s}_2$ *назовем взвешенный путь, имеющий наименьшую стоимость среди всех возможных матрично-взвешенных путей между последовательностями* $\boldsymbol{s}_1$ *и* $\boldsymbol{s}_2$*:*

$$\hat{\boldsymbol{\pi}} = \underset{\boldsymbol{\pi}}{argmin}\, Cost(\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{W}, \boldsymbol{\pi}). \tag{6}$$

Обозначим стоимость матрично-взвешенного выравнивающего пути между последовательностями $\mathbf{s}_1$ и $\mathbf{s}_2$ через $\rho(\mathbf{s}_1, \mathbf{s}_2, \mathbf{W}) = \mathrm{Cost}(\mathbf{s}_1, \mathbf{s}_2, \mathbf{W}, \hat{\boldsymbol{\pi}})$.

Вычисление его стоимости происходит с помощью еще одного модифицированного метода DTW — mwDTW (matrix-weighted DTW). Согласно этому методу необходимо построить новую матрицу $\boldsymbol{\Gamma}$, элементы которой определяются следующим образом:

$$\Gamma_{1j} = W_{1j}\Omega_{1j}, \quad \Gamma_{i1} = W_{i1}\Omega_{i1}, \quad i, j = 1, \ldots, n,$$

$$\Gamma_{ij} = W_{ij}\Omega_{ij} + \min(\Gamma_{i,j-1}, \Gamma_{i-1,j}, \Gamma_{i-1,j-1}), \quad i, j = 2, \ldots, n.$$

Элемент Γ_{ij} матрицы $\boldsymbol{\Gamma}$ равен стоимости матрично-взвешенного выравнивающего пути между последовательностями $\{s_{1a}\}_{a=1}^{i}$ и $\{s_{2a}\}_{a=1}^{j}$.

В качестве значения функции расстояния между двумя объектами выберем стоимость матрично-взвешенного выравнивающего пути между ними (6):

$$\rho(\mathbf{s}_1, \mathbf{s}_2, \mathbf{W}) = \Gamma_{nn}. \tag{7}$$

Заметим, что при использовании матрицы, состоящей из одних единиц, mwDTW переходит в обычный DTW, описание которого приведено в [15].

2.3 Вычисление параметров модели классификатора

2.3.1 Построение центроида

Пусть множество весов $\hat{\mathbf{W}}$ фиксировано. Построим множество центроидов $\mathbf{C}$.

2.3.2 Постановка задачи построения центроида

Определение 3. *Пусть $\mathfrak{D}_e$ — множество элементов из $\mathfrak{D}$, принадлежащих одному классу e из Y. Центроидом множества векторов $\mathfrak{D}_e = \{\boldsymbol{s}_i | y_i = e\}_{i=1}^m$ по расстоянию ρ назовем вектор $\boldsymbol{c}_e \in \mathbb{R}^n$ такой, что*

$$\boldsymbol{c}_e = \underset{\boldsymbol{c}\in\mathbb{R}^n}{argmin} \sum_{\boldsymbol{s}_i \in \mathfrak{D}_e} \rho(\boldsymbol{s}_i, \boldsymbol{c}),$$

где ρ — стоимость векторно- (матрично-) взвешенного пути наименьшей стоимости vwDTW (mvDTW).

Центроид найдем как решение оптимизационной задачи для vwDTW

$$\mathbf{c}_e = \underset{\mathbf{c}\in\mathbb{R}^n}{\operatorname{argmin}} \sum_{\mathbf{s}_i \in \mathfrak{D}_e} \sum_{(t,t')\in\hat{\boldsymbol{\pi}}_i} \mathbf{w}_e(t)\big(\mathbf{s}_i(t') - \mathbf{c}(t)\big)^2. \tag{8}$$

Или же для mwDTW:

$$\mathbf{c}_e = \underset{\mathbf{c}\in\mathbb{R}^n}{\operatorname{argmin}} \sum_{\mathbf{s}_i \in \mathfrak{D}_e} \sum_{(t,t')\in\hat{\boldsymbol{\pi}}_i} \mathbf{W}_e(t,t')\big(\mathbf{s}_i(t') - \mathbf{c}(t)\big)^2, \tag{9}$$

где $\hat{\boldsymbol{\pi}}_i$ — векторно- (матрично-) взвешенный выравнивающий путь между временными рядами $\mathbf{s}_i$ и $\mathbf{c}$.

2.3.3 Решение задачи нахождения центроида методом DBA

Теорема 1 [16]**.** *Пусть дано множество векторов $\mathfrak{D}_e = \{\boldsymbol{s}_i | y_i = e\}_{i=1}^m$ одного класса, начальное приближение центроида $\boldsymbol{c}_e$ и множество выравнивающих путей между каждым рядом и начальным приближением центроида $\{\tilde{\boldsymbol{\pi}}_i\}_{i=1}^m$. Тогда локальный минимум задачи оптимизации (8) при единичном векторе весов в (4) (функция расстояния DTW) достигается при*

$$\boldsymbol{c}_e(t) = \frac{1}{N} \sum_{\boldsymbol{s}_i \in \mathfrak{D}_e} \sum_{t':(t,t')\in\tilde{\boldsymbol{\pi}}_i} \boldsymbol{s}_i(t'), \tag{10}$$

$$N = \sum_{\boldsymbol{s}_i \in \mathfrak{D}_e} \sum_{t':(t,t')\in\tilde{\boldsymbol{\pi}}_i} 1.$$

Доказательство. *Для поиска центроида и решения задачи оптимизации (8) воспользуемся необходимым условием экстремума. Запишем*

частные производные функционала по $\boldsymbol{c}_e(t), t = 1, \ldots, T$, *и приравняем их к 0:*

$$\frac{\partial F(\boldsymbol{c}_e, \mathfrak{D}_e)}{\partial \boldsymbol{c}_e(t)} = \sum_{\boldsymbol{s}_i \in \mathfrak{D}_e} \sum_{t':(t,t') \in \tilde{\boldsymbol{\pi}}_i} 2\big(\boldsymbol{c}_e(t) - \boldsymbol{s}_i(t')\big) = 0. \tag{11}$$

Откуда и находим значение $\boldsymbol{c}_e(t), \quad t = 1, \ldots, n$*:*

$$\boldsymbol{c}_e(t) = \frac{1}{N} \sum_{\boldsymbol{s}_i \in \mathfrak{D}_e} \sum_{t':(t,t') \in \tilde{\boldsymbol{\pi}}_i} \boldsymbol{s}_i(t'). \tag{12}$$

Данный метод вычисления центроида приведен в [16] и называется методом DBA. Там же приведено и его доказательство. При нахождении нового центроида множество выравнивающих рядов меняется, данную процедуру нужно проводить несколько раз, пока центроид не стабилизируется. При замене единичного вектора весов в функции расстояния vwDTW на произвольный справедливо следующее

Следствие 1. *При использовании произвольного вектора весов центроида* $\boldsymbol{w}$ *(замене DTW на vwDTW с вектором весов* $\boldsymbol{w}$*) в задаче оптимизации (8) алгоритм DBA вычисления центроида находит локальный минимум при замене множества путей наименьшей стоимости* $\{\tilde{\boldsymbol{\pi}}\}_{i=1}^m$ *на множество взвешенных путей наименьшей стоимости* $\{\hat{\boldsymbol{\pi}}_i\}_{i=1}^m$.

Доказательство данного следствия повторяет доказательство теоремы 1 при замене множества путей наименьшей стоимости $\{\hat{\boldsymbol{\pi}}\}_{i=1}^m$ на множество взвешенных путей наименьшей стоимости $\{\boldsymbol{\pi}_i\}_{i=1}^m$.

Для функции расстояния mwDTW следствие сохраняет свою формулировку и доказательство при замене вектора весов центроида на его матрицу весов $\mathbf{W}$.

2.3.4 Оптимизация и ограничения вектора весов

Положим теперь множество центроидов $\mathbf{C}$ фиксированным. Каждому центроиду $\mathbf{c}_e$ из множества $\mathbf{C}$ поставлен в соответствие вектор неотрицательных весов $\mathbf{w}_e$, принадлежащий множеству $\hat{\mathbf{W}}$. Значения данного вектора весов выделяют наиболее типичные для класса участки центроида, сопоставляя им большие веса. Вычислим этот вектор, решая задачу оптимизации

$$\mathbf{w}_e = \underset{\mathbf{w}\in\mathbb{R}^n}{\operatorname{argmin}} \sum_{\mathbf{s}_i\in\mathfrak{D}_e} \sum_{(t,t')\in\boldsymbol{\pi}_i} \mathbf{w}(t)\big(\mathbf{s}_i(t') - \mathbf{c}_e(t)\big)^2. \tag{13}$$

При отсутствии ограничений на веса $\mathbf{w}_e$ минимум (13) достигается при $\mathbf{w}_e = \mathbf{0}$. Для того чтобы избежать такого тривиального решения, введем ограничения на сумму элементов вектора весов

$$\sum_{t=1}^{T} \mathbf{w}_e(t) = T.$$

Предположим, что при решении задачи (13) нашлось t, для которого выполняется

$$\sum_{\mathbf{s}_i\in\mathbf{D}} \sum_{t':(t,t')\in\boldsymbol{\pi}_i} \big(\mathbf{s}_i(t') - \mathbf{c}_e(t)\big)^2 = 0.$$

Для таких t элемент решения задачи оптимизации (13) $\mathbf{w}_e(t)$ примет большие значения, которые обеспечат выполнение ограничений на сумму элементов. Это приведет к локальному скоплению больших значений вектора весов, что сильно ухудшит дальнейшую интерпретацию вектора весов, а также сделает метод чувствительным к малым изменениям входных данных. Поэтому введем ограничения на элементы вектора весов сверху:

$$\mathbf{w}_e(t) \leq const, t \in \{1, ..., T\}.$$

Таким образом, исходная задача (13) примет следующий вид:

$$\mathbf{w}_e = \underset{\mathbf{w}\in\mathbb{R}^n}{\operatorname{argmin}} \sum_{\mathbf{s}_i\in\mathfrak{D}_e} \sum_{(t,t')\in\boldsymbol{\pi}_i} \mathbf{w}_e(t)\big(\mathbf{s}_i(t') - \mathbf{c}_e(t)\big)^2, \tag{14}$$

$$\sum_{t=1}^{T} \mathbf{w}_e(t) = T, \quad 0 \leq \mathbf{w}_e(t) \leq const, \quad t \in \{1, ..., T\},$$

где *const* — некоторая заданная константа.

2.3.5 Оптимизация матрицы весов

Ограничения матрицы весов Как и при оптимизации вектора весов положим множество центроидов $\mathbf{C}$ фиксированным. Каждому центроиду $\mathbf{c}_e$ некоторого класса e из множества $\mathbf{C}$ сопоставлена матрица неотрицательных весов $\mathbf{W}_e$, принадлежащая множеству $\hat{\mathbf{W}}$.

Используя те же соображения, что и для случая использования расстояния vwDTW, определим задачу нахождения матрицы весов центроида как задачу оптимизации с ограничениями при использовании расстояния mwDTW:

$$\mathbf{W}_e = \underset{\mathbf{W}\in\mathbb{R}^{n\times n}}{\operatorname{argmin}} \sum_{\mathbf{s}_i\in\mathfrak{D}_e} \sum_{(t,t')\in\boldsymbol{\pi}_i} \mathbf{W}_e(t,t')(\mathbf{s}_i(t') - \mathbf{c}_e(t))^2, \tag{15}$$

$$\sum_{t=1}^{T}\sum_{t'=1}^{T} \mathbf{W}_e(t,t') = T^2, \quad 0 \le \mathbf{W}_e(t) \le const, \quad t \in \{1,...,T\},$$

где *const* — некоторая заданная константа.

Сглаживание полученной матрицы Полученная матрица не является устойчивой к изменению входных данных: при использовании других временных рядов выравнивающие пути будут иметь другой вид, а значит, и решение задачи оптимизации будет другое. Более устойчивой матрица получится после процедуры сглаживания.

Будем говорить, что элемент матрицы $\mathbf{W}_e(t,t')$ содержится в множестве Φ, если существует временной ряд $\mathbf{s}_i \in \mathfrak{D}_e$ такой, что путь наименьшей матрично-взвешенной стоимости проходит через элемент $\Omega(t,t')$ в матрице Ω, построенной для временного ряда $\mathbf{s}_i$ и центроида.

При решении задачи оптимизации изменяться в меньшую сторону будут элементы $\{\mathbf{W}_e(t,t')\} \in \Phi$. При этом $\mathbf{W}_e(t,t') \notin \Phi$ достигнет своей верхней границы для выполнения ограничений, накладываемых на сумму элементов матрицы.

Выберем произвольный элемент матрицы весов $\mathbf{W}_e(t,t') \notin \Phi$. Вероятность того, что при добавлении нового временного ряда (например из тестовой выборки) выполнится $\mathbf{W}_e(t,t') \in \Phi$, выше, если среди ближайших к $\mathbf{W}_e(t,t')$ элементов в строке матрицы весов многие содержатся в множестве Φ. Тогда значение такого элемента должно быть похожим на значения соседних. Добьемся этого, выполнив сглаживание матрицы весов центроида:

$$\widetilde{\mathbf{W}}_e(t,t') = \frac{1}{2\delta}\sum_{k=-\delta}^{\delta} \mathbf{W}_e(t,t'+k),$$

где δ — величина окна сглаживания, а $\widetilde{\mathbf{W}}_e$ — искомая матрица весов центроида.

2.3.6 Задача оптимизации параметров модели

Задача оптимизации параметров модели сведена к комбинации задач оптимизации (8) и (14), (15) для vwDTW:

$$\mathbf{w}_e, \mathbf{c}_e = \underset{\mathbf{c},\mathbf{w}\in\mathbb{R}^n}{\operatorname{argmin}} \sum_{\mathbf{s}_i\in\mathfrak{D}_e} \sum_{(t,t')\in\boldsymbol{\pi}_i} \big(\mathbf{w}(t)(\mathbf{s}_i(t') - \mathbf{c}(t))^2\big), \quad e = 1, \ldots, E, \tag{16}$$

$$\sum_{t=1}^{T} \mathbf{w}_e(t) = T, \quad 0 \le \mathbf{w}_e(t) \le const, \quad t \in \{1, ..., T\}.$$

Или же для mwDTW:

$$\mathbf{W}_e, \mathbf{c}_e = \underset{\mathbf{W}\in\mathbb{R}^n\times\mathbb{R}^n}{\operatorname{argmin}} \sum_{\mathbf{s}_i\in\mathfrak{D}_e} \sum_{(t,t')\in\boldsymbol{\pi}_i} \big(\mathbf{W}_e(t,t')(\mathbf{s}_i(t') - \mathbf{c}_e(t))^2\big), \quad e = 1, \ldots, E, \tag{17}$$

$$\sum_{t=1}^{T}\sum_{t'=1}^{T} \mathbf{W}_e(t,t') = T^2, \quad 0 \le \mathbf{W}_e(t) \le const, \quad t \in \{1, ..., T\}.$$

Эту задачу будем решать, вычисляя сначала множество центроидов $\mathbf{C}$ при фиксированном начальном приближении множества весов центроидов $\hat{\mathbf{W}}$, а затем вычисляя множество весов центроидов $\hat{\mathbf{W}}$ при фиксированном множестве центроидов $\mathbf{C}$. Таким образом, алгоритм вычисления параметров модели будет иметь следующий вид:

Шаг 1. Начальное приближение вектора весов центроида:

$$\mathbf{w}_e = \mathbf{1}, \quad e = 1, \ldots, E.$$

Или же начальное приближение матрицы весов центроида:

$$\mathbf{W}_e = \mathbf{1}, \quad e = 1, \ldots, E.$$

Шаг 2. Начальное приближение центроида класса — произвольный элемент класса:

$$\mathbf{c}_e = \mathbf{s}_j \in \mathfrak{D}_e, \quad e = 1, \ldots, E.$$

Шаг 4. Вычисление центроида при фиксированном векторе (матрице) весов центроида как решение задачи оптимизации (8).

Шаг 3. Вычисление вектора (матрицы) весов центроида при фиксированном центроиде как решение задачи оптимизации (14), (15).

2.4 Вычислительный эксперимент раздела 2

Для проверки свойств введенной функции расстояния, а также выбранной модели был проведен вычислительный эксперимент на реальных и синтетических данных. Свойства функций расстояния vwDTW и mwDTW для наглядности продемонстрированы на синтетических данных, представляющих собой смещенные и линейно деформированные временные ряды аналитических функций: $sinx, \sqrt{x}, x^2$, — длиной 100 точек. В выборке находилось 100 временных рядов каждого класса: 50 в обучающей и 50 в контрольной. Обозначим классы как $1 - sinx, \quad 2 - \sqrt{x}, \quad 3 - x^2$ соответственно.

Пример такой выборки и результаты нахождения параметров модели по обучающей выборке показаны на рис. 1. По оси абсцисс отложены значения времени, а по оси ординат — значения временного ряда. На левом верхнем графике приведены примеры аналитических функций, используемых в создании синтетической выборки временных рядов. На правом верхнем графике изображены центроид (нижний временной ряд) и вектор весов (верхний временной ряд) для класса 1. Аналогично на нижних графиках показаны результаты для классов 2 и 3.

Векторы весов описывают наиболее информативные участки центроида. Так, для центроида класса 1 ($sinx$) наиболее информативными оказались минимумы и максимумы, в отличие от точек перегиба.

Для сравнения свойств полученной функции расстояния с функцией расстояния DTW были посчитаны расстояния до центроидов для всех временных рядов контрольной выборки с помощью DTW и vwDTW, после чего производилась классификация. Каждому временному ряду контрольной выборки ставилась в соответствие метка того класса, расстояние до центроида которого было минимальным. Результат классификации с помощью функции расстояния vwDTW — 97%, а для функции расстояния DTW — 84%, что на 15% меньше. Построенная в работе функция расстояния лучше разделила временные ряды разных классов, сгруппировала их вокруг соответствующих центроидов.

Также были построены матрицы весов центроида на синтетических данных. Вследствие хорошей интерпретируемости, в работе приведено визуальное отображение матрицы весов центроида только для класса 1. На рис. 2 сверху слева изображены примеры временных рядов. Сверху справа — множество выравнивающих путей по матрице $\mathbf{\Omega}$. Нижний левый рисунок — матрица весов центроида до сглаживания. Правый ле-

вый рисунок — после сглаживания. Синие цвета соответствуют маленьким значениям элементов матрицы, а красные — большим. Хорошо просматривается периодичность в матрице, напоминающая периодичность вектора весов. Предполагается, что функция mwDTW будет лучше разделять классы, которые сильно различаются между собой, так как матрица весов центроида учитывает и среднее отклонение выравнивающего пути от диагонали в матрице $\mathbf{\Omega}$. При отклонениях пути сильнее типичного для данного класса, элементам пути будут приписываться большие веса, что видно из структуры матрицы весов центроида.

Использование функции mwDTW для классификации синтетических временных рядов улучшает классификацию по сравнению с DTW. Этот результат аналогичен случаю использования vwDTW — 97%.

Для демонстрации работы предложенной модели на реальных данных и ее сравнения с базовой моделью классификации были использованы данные акселерометра мобильного телефона. Вследствие большой вычислительной сложности они сравнивались на данных, представляющих собой 600 временных рядов длиной 200 точек, каждый из которых представляет собой абсолютные значения ускорения мобильного телефона, объединяя три временных ряда: временной ряд ускорения по оси X (200 измерений), оси Y (200 измерений) и оси Z (200 измерений). Выделено шесть типов физической активности: ходьба, бег, сидение, стояние, подъем, спуск. Временные ряды записывались акселерометром, который находился в кармане у человека, выполняющего один из типов физической активности, после чего разделялись на 10-секундные сегменты. Примеры таких временных рядов приведены на рис. 3.

Данные разделялись на обучающую и контрольную выборку. В обучающую выборку входило по 70 временных рядов каждого вида физической активности, а в контрольную — по 30 временных рядов. Производилась классификации методом k ближайших соседей, использующим построенную матрицу попарных расстояний. Осуществлялся контроль качества при помощи кросс-валидации. В табл. 1 приведены результаты классификации при использовании новой модели, использующей функцию vwDTW, и базовой модели классификации.

Качество классификации базовой модели ухудшилось в сравнении с вычислительным экспериментом в работе [15], так как теперь используются абсолютные значения ускорения, а не последовательно соединенные временные ряды ускорения вдоль трех координат. Использование функции расстояния vwDTW улучшает классификацию для всех клас-

сов физической активности, повышая общий уровень классификации на 9%.

Таблица 1: Сравнение эффективности предложенной (vwDTW) и базовой модели классификации на данных [17]

Модель	Бег	Ходьба	Вверх	Вниз	Сидение	Стояние	Общее
vwDTW. Точность по критерию скользящего контроля, %	97	95	79	75	95	95	89
DTW [15]. Точность по критерию скользящего контроля, %	95	92	60	60	85	90	80

Полученные параметры модели для реальных данных изображены на рис. 4. Для класса бег, например, хорошо просматривается периодичность временного ряда как для центроида, так и для вектора весов центроида.

Для сравнения модели, использующей mwDTW, с базовой моделью классификации были использованы данные, представляющие собой 600 временных рядов длиной 100 точек. Временные ряды такой длины выбраны для разумного ограничения времени работы методов численной оптимизации. В остальном этот эксперимент повторяет тот, что выполнен для сравнения базовой модели и модели, использующей vwDTW. Результаты классификации приведены в табл. 2. Видно, что длина временных рядов, а следовательно, и количество периодов временного ряда, существенно влияют на качество классификации. Качество базовой модели сильно упало по сравнению с временными рядами длиной 200 точек. При этом использование mwDTW улучшает классификацию на 13%.

Таблица 2: Сравнение эффективности предложенной (mwDTW) и базовой модели классификации и алгоритма разделяющей классификации на данных [17]

Модель	Бег	Ходьба	Вверх	Вниз	Сидение	Стояние	Общее
mwDTW. Точность по критерию скользящего контроля, %	95	95	78	76	90	90	87
DTW [15]. Точность по критерию скользящего контроля, %	88	83	55	60	82	80	74

3 Выравнивание для случая непрерывных объектов

3.1 Задача представления временного ряда непрерывным аналогом

В дискретном случае временной ряд $\mathbf{s}$ представляет собой упорядоченную во времени последовательность измерений какой-либо величины. Введем определение непрерывного временного ряда:

Определение 1 *Непрерывный временной ряд, определенный на участке времени* $\widehat{T} = [0; T]$, *— непрерывная функция* $s^c(t) : \widehat{T} \to \mathbb{R}$.

Обозначим через $\mathcal{S}$ пространство всевозможных дискретных временных рядов. Через $\mathcal{S}_c$ — пространство всевозможных непрерывных временных рядов. Для получения непрерывного аналога временного ряда каждому объекту $s \in \mathcal{S}$ ставится в соответствие объект $s(t) \in \mathcal{S}_c$. Необходимо построить отображение $f : \mathcal{S} \to \mathcal{S}_c$.

В качестве непрерывного аналога дискретного временного ряда в работе используется его аппроксимация параметрическими функциями. Алгоритм ее построения выполняется в три шага: выбор подходящего пространства параметрических функций, подбор оптимальных параметров для фиксированного временного ряда, использование оптимальных параметров для построения конкретной аппроксимации.

3.1.1 Аппроксимация функции кубическими сплайнами

В работе будет использована аппроксимация кубическими сплайнами (ссылка на работу).

Пусть дано множество точек на плоскости $\{(x_i, y_i)\}_{i=1}^n$, порожденных произвольной непрерывной и гладкой функцией $f(x)$. Требуется построить аппроксимацию этой функции. Интерполяция функции между двумя соседними точками (x_i, y_i) и (x_{i+1}, y_{i+1}) производится при помощи полиномов третьей степени. При этом в заданных точках совпадают как значения соседних полиномов так и значения их производных и вторых производных. Обозначим интерполяцию функции $f(x)$ через $\widehat{f}(x)$. А через $\widehat{f}_i(x)$ — интерполяцию $\widehat{f}(x)$ при $x_i \le x \le x_{i+1}$. Тогда

$$\widehat{f}(x) = \widehat{f}_i(x), \quad x_i \le x \le x_{i+1},$$

$$\widehat{f}_i(x) = a_i(x-x_i)^3 + b_i(x-x_i)^2 + c_i(x-x_i) + d_i, \quad x_i \le x \le x_{i+1}, \quad i \in \{1, ..., n-1\},$$

Должно выполняться следующее:

$$\widehat{f}_i(x_{i+1}) = \widehat{f}_{i+1}(x_{i+1}), \quad i \in \{1, ..., n-2\},$$
$$\widehat{f}'_i(x_{i+1}) = \widehat{f}'_{i+1}(x_{i+1}), \quad i \in \{1, ..., n-2\},$$
$$\widehat{f}''_i(x_{i+1}) = \widehat{f}''_{i+1}(x_{i+1}), \quad i \in \{1, ..., n-2\},$$

где n — количество заданных точек.

Заданное множество точек $\{(x_i, y_i)\}_{i=1}^n$ будем называть узлами сплайна. В роли узлов сплайна в задаче аппроксимации временного ряда выступают временные отсчёты и измерения, им соответствующие.

3.2 Расстояние DTW и выравнивающий путь между непрерывными временными рядами

3.2.1 Базовые определения

В качестве функции расстояния между временными рядами в этой работе выбрана DTW. Эта функция определена между дискретными временными рядами. Для работы в непрерывном случае требуются определения стоимости пути наименьшей стоимости, выравнивающего пути и расстояния DTW для непрерывного случая.

Предположим, что имеется два дискретных временных ряда $\mathbf{s}_1$ и $\mathbf{s}_2$, а также их непрерывные аналоги, заданные непрерывными функциями: $s_1^c(t_1), \quad t_1 \in [0;T] \quad$ и $\quad s_2^c(t_2), \quad t_2 \in [0;T] \quad$. Также предположим, что обе функции кусочно-гладкие, и для обеих выполняется условие Липшица с константами Липшица L_1, L_2.

$$|s_1(t_1) - s_1(t_1')| \leq L_1|t_1 - t_1'|, \quad s_1 \in C^1_{[0;T]},$$

$$|s_2(t_2) - s_2(t_2')| \leq L_2|t_2 - t_2'|, \quad s_2 \in C^1_{[0;T]}.$$

Определение 1 (дискретный случай): *путь* $\boldsymbol{\pi}$ *между дискретными временными рядами* $\boldsymbol{s}_1$ *и* $\boldsymbol{s}_2$ *— упорядоченное множество пар индексов:*

$$\boldsymbol{\pi} = \{\pi_r\} = \{(i_r, j_r)\}, \quad r = 1, \ldots, R, \quad i, j \in \{1, ..., n\},$$

удовлетворяющее условиям непрерывности, монотонности и граничным условиям:

$$\pi_r = (p_1, p_2), \quad \pi_{r-1} = (q_1, q_2), \quad r = 2, ..., R, \quad \Rightarrow \quad p_1 - q_1 \leq 1, \quad p_2 - q_2 \leq 1,$$

$$\pi_r = (p_1, p_2), \quad \pi_{r-1} = (q_1, q_2), \quad r = 2, ..., R, \quad \Rightarrow \quad p_1 - q_1 \geq 1, \quad p_2 - q_2 \geq 1,$$

$$\pi_1 = (1, 1), \quad \pi_R = (n, n).$$

Определение 1 (непрерывный случай): *путь* π^c *между двумя непрерывными временными рядами — монотонно возрастающая, непрерывная функция* $\pi^c : t_1 \to t_2$ *, удовлетворяющая начальным условиям:*

$$\pi^c \in C_{[0;T]},$$
$$t_1 > t_1' \Rightarrow \pi^c(t_1) > \pi^c(t_1'),$$
$$\pi^c(0) = 0, \quad \pi^c(T_1) = T_2.$$

Определение 2 (дискретный случай): *Стоимость* $Cost(\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{\pi})$ *пути* $\boldsymbol{\pi}$ *длины* R *между дискретными временными рядами* $\boldsymbol{s}_1$ *и* $\boldsymbol{s}_2$ *:*

$$\mathrm{Cost}(\mathbf{s}_1, \mathbf{s}_2, \boldsymbol{\pi}) = \frac{1}{R} \sum_{(i,j) \in \boldsymbol{\pi}} |\mathbf{s}_1(i) - \mathbf{s}_2(j)|.$$

Определение 2 (непрерывный случай): *Стоимость* $Cost(s_1^c(t_1), s_2^c(t_2), \pi^c)$ *пути* π^c *между непрерывными временными рядами* $s_1^c(t_1)$ *и* $s_2^c(t_2)$ *:*

$$\mathrm{Cost}(s_1^c(t_1), s_2^c(t_2), \pi^c) = \frac{1}{L}\int_{t_1} |s_1^c(t_1) - s_2^c(\pi^c(t_1))| dt_1,$$

где L — длина кривой, задающейся графиком функции $\pi^c(t), \quad t \in [0, T]$.

Определение 3 (дискретный случай):*Путь наименьшей стоимости (выравнивающим путем)* $\widehat{\boldsymbol{\pi}}$ *между дискретными временными рядами* $\boldsymbol{s}_1$ *и* $\boldsymbol{s}_2$ *— путь, имеющий наименьшую стоимость среди всех возможных путей:*

$$\widehat{\boldsymbol{\pi}} = \underset{\boldsymbol{\pi}}{\operatorname{argmin}} \mathrm{Cost}(\mathbf{s}_1, \mathbf{s}_2, \boldsymbol{\pi}).$$

Определение 3 (непрерывный случай): *Путь наименьшей стоимости (выравнивающий путь)* $\widehat{\pi}^c$ *между непрерывными временными рядами* $s_1^c(t_1)$ *и* $s_2^c(t_2)$ *— функция* $\widehat{\pi}^c$*, для которой значение интеграла из определения 2 для непрерывного случая является наименьшим:*

$$\widehat{\pi}^c = \underset{\pi^c}{\operatorname{argmin}} \mathrm{Cost}(s_1^c(t_1), s_2^c(t_2), \pi^c).$$

Определение 4 (дискретный случай): *Стоимость пути наименьшей стоимости, или расстояние DTW между дискретными временными рядами:*

$$DTW(\mathbf{s}_1, \mathbf{s}_2) = \mathrm{Cost}(\mathbf{s}_1, \mathbf{s}_2, \widehat{\boldsymbol{\pi}}).$$

Определение 4 (непрерывныйслучай): *Стоимость пути наименьшей стоимости, или расстояние DTW между непрерывными временными рядами:*

$$DTW(s_1^c(t_1), s_2^c(t_2)) = \mathrm{Cost}(s_1^c(t_1), s_2^c(t_2), \widehat{\pi}^c).$$

3.2.2 Свойства выравнивающего пути и его стоимости

Сформулируем две леммы, справедливые для случая непрерывных временных рядов.

Лемма 1. *Предположим, что $s_1(t)$ и $s_2(t)$ – два временных ряда, $\widehat{\pi}^c : t_1 \to t_2$ — выравнивающий путь между ними. При малых изменениях пути, его стоимость изменяется слабо, то есть:*

$$\| \widehat{\pi}^c - \pi^c \|_C \leq \epsilon \quad \Rightarrow \quad |Cost(s_1, s_2, \widehat{\pi}^c) - Cost(s_1, s_2, \pi^c)| \leq \epsilon T L,$$

где L — константа Липшица для $s_1(t)$ и $s_2(t)$, T — граница области определения временного ряда, $\epsilon > 0$.

Доказательство.

$$|Cost(s_1, s_2, \widehat{\pi}^c) - Cost(s_1, s_2, \pi^c)| = |\int_{t_1} |s_1(t_1) - s_2(\widehat{\pi}^c(t_1))| dt_1 - \int_{t_1} |s_1(t_1) - s_2(\pi^c(t_1))| dt_1| \leq$$

$$\int_{t_1} |s_2(\widehat{\pi}^c(t_1)) - s_2(\pi^c(t_1))| dt_1 \quad \leq \quad \int_{t_1} L |\widehat{\pi}^c(t_1) - \pi^c(t_1)| dt_1 \quad \leq$$

$$\int_{t_1} L \max_{t_1 \in [0,T]} (\widehat{\pi}^c(t_1) - \pi^c(t_1)) dt_1 \quad \leq \quad \int_{t_1} L \| \widehat{\pi}^c - \pi^c \|_C \, dt_1 \quad \leq \quad TL\epsilon. \quad \square$$

Лемма 2. *Предположим, что $s_1(t)$ и $s_2(t)$ – два временных ряда, $\widehat{\pi}^c : t_1 \to t_2$ — выравнивающий путь между ними. При малых изменениях одного из временных рядов, стоимость пути изменяется слабо, то есть:*

$$\| \widehat{s}_2 - s_2 \|_C \leq \epsilon \quad \Rightarrow \quad |Cost(s_1, \widehat{s}_2, \widehat{\pi}^c) - Cost(s_1, s_2, \widehat{\pi}^c)| \leq \epsilon T L,$$

где L — константа Липшица для $s_1(t)$ и $s_2(t)$, T — граница области определения временного ряда, $\epsilon > 0$.

Доказательство.

$$|Cost(s_1, \widehat{s}_2, \widehat{\pi}^c) - Cost(s_1, s_2, \widehat{\pi}^c)| = |\int_{t_1} |s_1(t_1) - \widehat{s}_2(\widehat{\pi}^c(t_1))| dt_1 - \int_{t_1} |s_1(t_1) - s_2(\widehat{\pi}^c(t_1))| dt_1| \leq$$

$$\int_{t_1} |\widehat{s}_2(\widehat{\pi}^c(t_1)) - s_2(\widehat{\pi}^c(t_1))| dt_1 \quad \leq \quad \int_{t_1} \max_{t_1 \in [0,T]} (\widehat{s}_2(\widehat{\pi}^c(t_1)) - s_2(\widehat{\pi}^c(t_1))) dt_1 \quad \leq$$

$$\int_{t_1} \| \widehat{s}_2 - s_2 \|_C \, dt_1 \quad \leq \quad T\epsilon. \quad \square$$

Эти две леммы демонстрируют свойство устойчивости стоимости пути к незначительным изменениям в начальных данных и в выравнивающем пути. Также выдвигается предположение об устойчивости выравнивающего пути к небольшому изменению начальных данных, то есть:

Предположение 1. *Предположим, что $s_1(t)$ и $s_2(t)$ – два временных ряда. Предположим, что $\widehat{s}_2(t)$ — временной ряд, слабо отличающийся от $s_2(t)$. Тогда*

$$\forall \epsilon_1 > 0 \quad \exists \epsilon_2(\epsilon_1), \quad \forall \widehat{s_2}(t): \quad \| \widehat{s}_2(t) - s_2(t) \|_C \leq \epsilon_2 \quad \mapsto \quad \| \pi^c - \widehat{\pi^c} \|_C \quad \leq \quad \epsilon_1,$$

где π^c и $\widehat{\pi^c}$ — выравнивающие пути между $s_1(t), s_2(t)$ и $s_1(t), \widehat{s_2}(t)$ соответственно.

3.3 Постановка задачи вычисления пути наименьшей стоимости

Пусть $s_1(t_1)$ и $s_2(t_2)$ — временные ряды. Требуется построить функцию $\widehat{\pi}^c : t_1 \to t_2$, являющуюся решением оптимизационной задачи из определения 3:

$$\widehat{\pi}^c = \underset{\pi^c}{\operatorname{argmin}} \operatorname{Cost}(s_1^c(t_1), s_2^c(t_2), \pi^c).$$

В случае дискретных временных рядов алгоритм нахождения выравнивающего пути связан с перебором конечного множества различных путей между временными рядами. В непрерывном случае воспользоваться перебором невозможно, так как множеством путей является множеством монотонных и непрерывных функций, удовлетворяющих соответствующим ограничениям из определения 1.

Решить описанную выше задачу оптимизации напрямую в пространстве функций нельзя. Предлагается ограничить множество, в котором выполняется поиск $\widehat{\pi}^c$. Выбирается класс параметрических функций, среди которых и будет осуществляться поиск. Каждая функция в таком множестве однозначно определяется набором параметров. А значит, поиск функции в задаче оптимизации сводится к поиску подходящих параметров. Эта задача может быть решена с помощью численных методов оптимизации.

Найденная функция $\widehat{\widehat{\pi}}^c$ является аппроксимацией выравнивающего пути$\widehat{\pi}^c$ между временными рядами. Если аппроксимация хорошая, то есть выполняется:

$$\| \widehat{\widehat{\pi}}^c - \widehat{\pi}^c \|_C \leq \epsilon,$$

то согласно Лемме 1 стоимость такого пути не будет сильно отличаться от стоимости выравнивающего пути, а значит, можно использовать такую аппроксимацию.

Задача поиска сведена к задаче оптимизации по множеству параметров, задающих аппроксимацию выравнивающего пути:

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmin}}\, Cost(s_1, s_2, \theta) = \underset{\theta}{\operatorname{argmin}} \int_{t_1} |s_1(t_1) - s_2(F(\theta)(t_1))| dt_1,$$

где $F(\theta)$ является отображением из пространства параметров в выбранное ранее пространство параметрических функций.

Как и в задаче аппроксимации временного ряда, для решения задачи аппроксимации пути наименьшей стоимости используется аппроксимация функции кубическими сплайнами. В роли параметров аппроксимации выступают узлы сплайна. Могут меняться как их координаты, так и их количество.

Предположим, что количество узлов N задано. Также известны их координаты по оси t_1. В таком случае параметрами аппроксимации выравнивающего пути будем считать координаты узлов сплайна по оси t_2, то есть $\theta = \{t_{2i}\}_{i=1}^{N}$. Свойство непрерывности для аппроксимации пути выполняется. Осталось наложить ограничения на параметры модели для выполнения граничных условий и свойства монотонности. Для этого примем

$$t_{21} = t_{11} \quad t_{2N} = t_{1N},$$

при этом

$$t_{2(i)} \leq t_{2(i+1)}, \quad i \in \{1, ..., N-1\}.$$

При таком выборе множества параметрических функций выполнена следующая

Лемма 3 *При небольшом изменении вектора параметров, задающих аппроксимацию выравнивающего пути, стоимость пути меняется слабо:*

$$\| \widehat{\theta} - \theta \|_2 \leq \epsilon \Rightarrow |Cost(s_1, \widehat{s}_2, F(\widehat{\theta})) - Cost(s_1, s_2, F(\theta))| \leq \delta.$$

Доказательство: При небольшом изменении координаты y_i узла i сплайна путь между временными рядами меняется несильно по норме $\| \cdot \|_C$. Дальнейшее доказательство следует из леммы 1. $\square$

3.4 Вычислительный эксперимент раздела 3

В вычислительном эксперименте исследовались свойства полученной функции расстояния, устойчивости метода к различным изменениям во входных данных. Использовались временные ряды акселерометра мобильного телефона. Временные ряды акселерометра представляли собой 600 временных рядов длиной 200 точек, каждый из которых представлял абсолютное значение ускорения, измерявшееся по трём координатам: временной ряд ускорения по оси X (200 измерений), оси Y (200 измерений) и оси Z (200 измерений). Выделено шесть типов физической активности: ходьба, бег, сидение, стояние, подъем, спуск по 10 временных рядов для каждого класса.

В начале эксперимента данные разделялены на 6 подмножеств, соответствующих классам физической активности. Для каждого подмножества был вычислен центроид класса, а потом для каждого временного ряда и центроида построен непрерывный аналог. Было выбрано количество узлов сплайна, задающее точное и, при этом, максимально быстрое вычисление расстояния, а параллельно с этим выявлена зависимость точности вычисления расстояния и времени его вычисления от количества узлов.

Итак, исходные данные представлены дискретными временными рядами, то есть упорядоченным множеством измерений $\{s_i)\}_{i=1}^n$. Для аппроксимации или же интерполяции этого ряда необходимо представить его в виде $\{(x_i, y_i)\}_{i=1}^n$. Положим $x_i = i, \quad y_i = s_i, \quad i = 1, \ldots, n$. В данном случае в качестве x_i взяты временные отсчеты, располагающиеся на временной оси на равных интервалах, что обусловлено типом записи временных рядов. При использовании других временных рядов, где известно точное время записи значения, можно изменить значения x_i.

Временные ряды интерполировались сплайнами третьей степени. Пример такой интерполяции представлен на рисунке 5. Точками обозначен временной ряд, а непрерывной линией — его интерполяция. Видно, что она является гладкой.

Если известен способ аппроксимации или же интерполяции временного ряда, который для конкретной задачи имеет высокую точность, можно с легкостью использовать его. Теория не накладывает никаких ограничений на использование других способов аппроксимации.

Такая же аппроксимация использовалась и для приближения пути наименьшей стоимости. В качестве параметров модели в работе исполь-

зованы координаты узлов сплайна по одной из осей. При этом координаты узлов по другой оси фиксированы. Гиперпараметром этой аппроксимации является количество узлов сплайна N. При малых N траектория пути не может подстроиться под форму временного ряда, и расстояние будет завышенным. Предполагается, что при увеличении N расстояние будет сходиться к истинному. Это подтверждается и вычислительными экспериментами, результат которых показан на рисунке 6.

Для исследования свойств разделяющей способности введенной функции расстояния между временными рядами исследовались матрицы попарных расстояний между временными рядами и центроидами классов. Эксперимент проводился на временных рядах акселерометра мобильного телефона.

В таблице 3 приведены средние значения функции расстояния между рядами класса и центроидами различных классов.

Таблица 3: Средние расстояния между объектами различных классов и центроидами этих классов для непрерывного случая.

	Бег	Ходьба	Вверх	Вниз	Сидение	Стояние
Бег	**693**	803	811	733	1165	1143
Ходьба	676	**498**	696	610	946	927
Вверх	714	739	**696**	701	1038	1021
Вниз	591	601	653	**464**	836	804
Сидение	516	465	434	400	**6**	42
Стояние	508	441	454	366	105	**79**

Из таблицы видно, что среднее расстояние от каждого класса минимально при использовании центроида именно этого класса. Данное свойство является важным для функций расстояния, так как демонстрирует разделяющую способность функции.

4 Заключение

В работе описан новый подход к работе с центроидами временных рядов и непрерывными объектами, построена модель, использующая веса и матрицы центроидов, и показаны ее преимущества перед моделью, описанной в работе [15]. Исследованы свойства моделей, разделяющая и объединяющая способности функций расстояния. Продемонстрировано влияние структуры, длины и физического смысла временных рядов на результаты классификации. В последующих работах данный подход будет совершенствоваться. Предполагается ускорить данный алгоритм за счет эффективного нахождения пути наименьшей стоимости, его аппроксимации. Такой подход применим и ко всем типам временных рядов, где возможна нелинейная деформация как по оси времени, так и по оси значений временного ряда.

Список литературы

[1] *Кузнецов М. П., Ивкин Н. П.* Алгоритм классификации временных рядов акселерометра по комбинированному признаковому описанию // Машинное обучение и анализ данных, 2015, Т. 1. Вып . 13. С. 1471–1483.

[2] *Попова М. С., Стрижов В. В.* Выбор оптимальной модели классификации физической активности по измерениям акселерометра // Информатика и ее применения, 2015, Т. 9. Вып . 1. С. 79–89.

[3] *Ignatov A. D., Strijov V. V.* Human activity types recognition using quasiperiodic sets of time series collected from a single tri-axial accelerometer // Multimedia tools and applications, 2015, Springer US, P. 1–14.

[4] *Faloutsos C., Ranganathan M., Manolopoulos Y.* Fast Subsequence Matching in Time-Series Databases // SIGMOD International Conference on Management of Data, 1994, Minneapolis, ACM, P. 419–429.

[5] *Berndt D. J., Clifford J.* Using dynamic time warping to find patterns in time series // Workshop on Knowledge Discovery in Databases, at

the 12th Int'l Conference on Artificial Intelligence, 1994, Seattle, WA, P. 359–370.

[6] *Keogh E. J., Ratanamahatana C. A.* Exact indexing of dynamic time warping // Knowl. Inf. Syst., 2005, Vol. 7. No. 3. P. 358–386.

[7] *Vlachos M., Gunopulos D., Kollios G.* Discovering similar multidimensional trajectories // IEEE International Conference on Data Engineering, 2002, San Jose, IEEE Computer Society, P. 673–684.

[8] *Chen L., Ng R. T.* On the marriage of lp-norms and edit distance // Very Large Data Bases (VLDB), 2004, Toronto, Morgan Kaufmann, P. 792–803

[9] *Chen L., Özsu M. T., Oria V.* Robust and fast similarity search for moving object trajectories // ACM International Conference on Management of Data (SIGMOD), 2005, Baltimore, ACM, P. 491–502.

[10] *Frentzos E., Gratsias K., Theodoridis Y.* Index-based most similar trajectory search // IEEE International Conference on Data Engineering (ICDE), 2007, Istanbul, IEEE Computer Society, P. 816–825.

[11] *Morse M. D., Patel J. M.* An efficient and accurate method for evaluating time series similarity // ACM International Conference on Management of Data (SIGMOD), 2007, Beijing, ACM, P. 569–580.

[12] *Chen Y., Nascimento M. A., Ooi B. C., Tung A. K. H.* SpADe: On Shape-based Pattern Detection in Streaming Time Series // IEEE International Conference on Data Engineering (ICDE), 2007, Istanbul, IEEE Computer Society, P. 786–795.

[13] *Keogh E. J., Pazzani M. J.* Scaling up Dynamic Time Warping to Massive Datasets // Principles of Data Mining and Knowledge Discovery, 1999, Prague, Springer Berlin Heidelberg, P. 1–11.

[14] *Salvador S., Chan P.* Fastdtw: Toward accurate dynamic time warping in linear time and space // Workshop on Mining Temporal and Sequential Data, 2004, Seattle, P. 70–80.

[15] *Гончаров А. В., Попова М. С., Стрижов В. В.* Метрическая классификация временных рядов с выравниванием относительно центроидов классов // Системы и средства информатики, 2015, Т. 25. Вып. 4. С. 52–64.

[16] *Petitjean F., Forestier G., Webb G. I., Nicholson A. E., Chen Y., Keogh E.* Dynamic Time Warping Averaging of Time Series allows Faster and more Accurate Classification // IEEE International Conference on Data Engineering (ICDE), 2014, Chicago, IEEE Computer Society, P. 470–479.

[17] Data from accelerometer. Available at: http://sourceforge.net/p/mlalgorithms/ TSLearning/data/preprocessed_large.csv (accessed November 15, 2015).

[18] *Carl de Boor.* A Practical Guide to Splines // Springer-Verlag. 1978. P. 113-114.

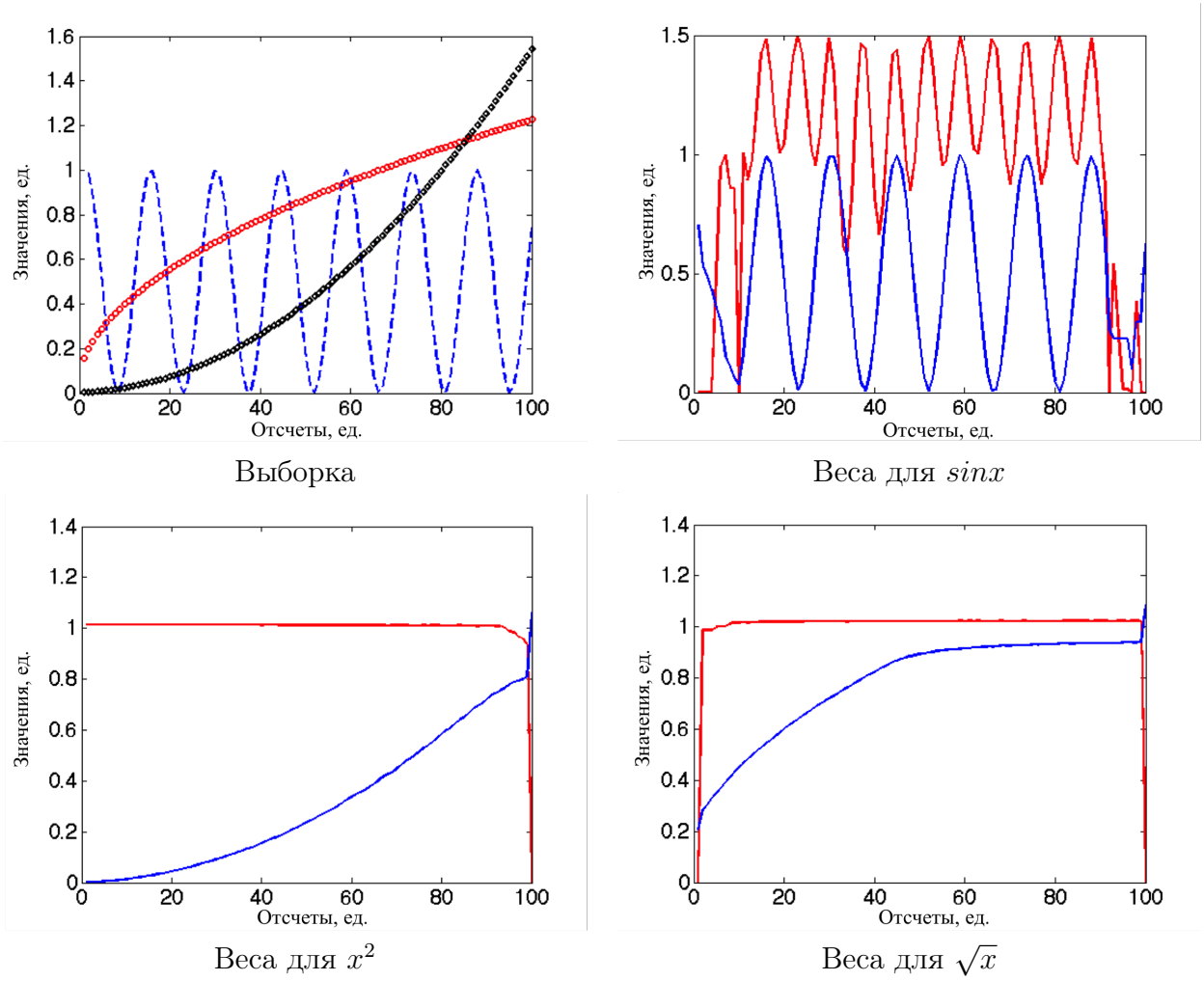
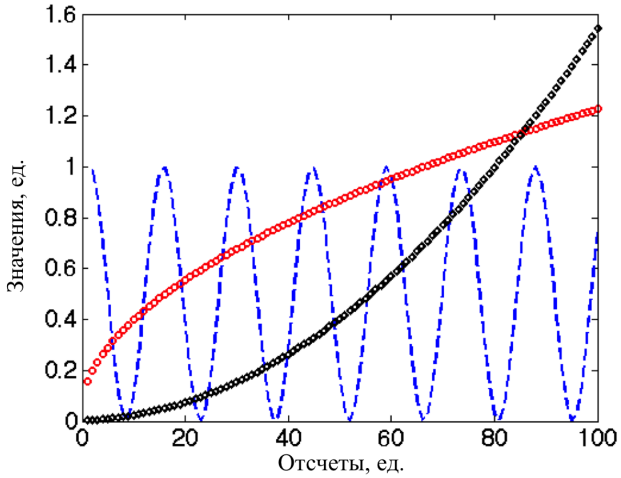
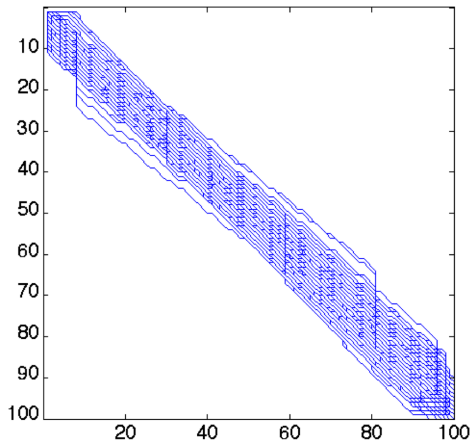


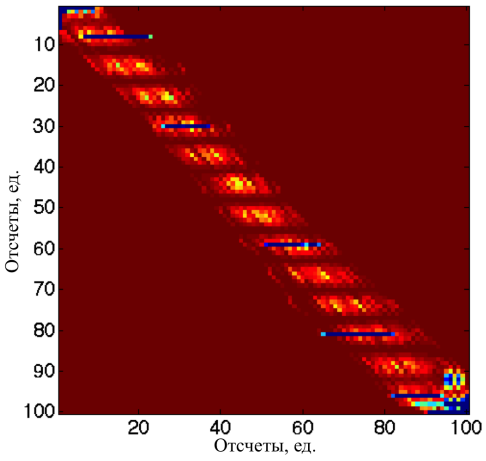
Рис. 1: Примеры синтетических временных рядов аналитических функций (выборка), результаты построения центроида и вектора весов для синтетических данных



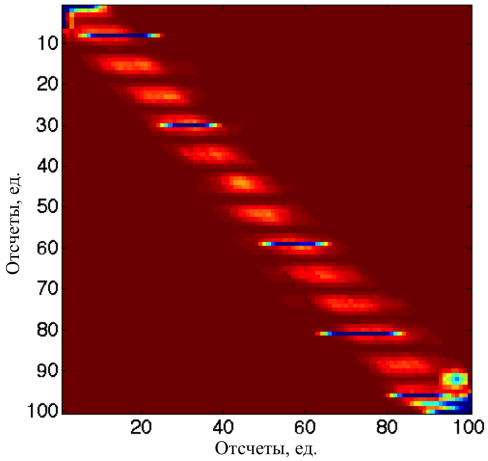
Выборка



Множество выравнивающих путей



Начальная матрица весов



Сглаженная матрица весов

Рис. 2: Примеры синтетических временных рядов аналитических функций (выборка), результат построения матрицы весов центроида класса sin(x)

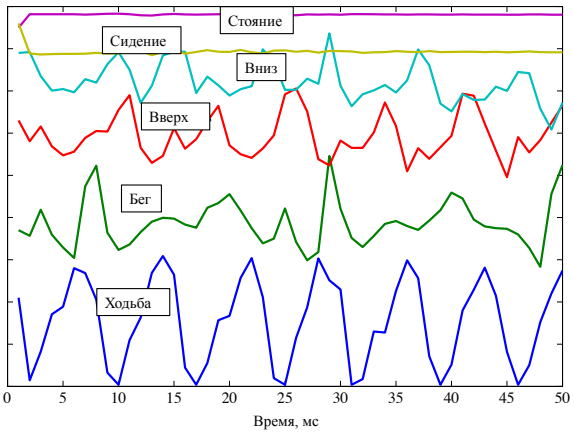


Рис. 3: Примеры временных рядов измерений акселерометра для разных видов физической активности

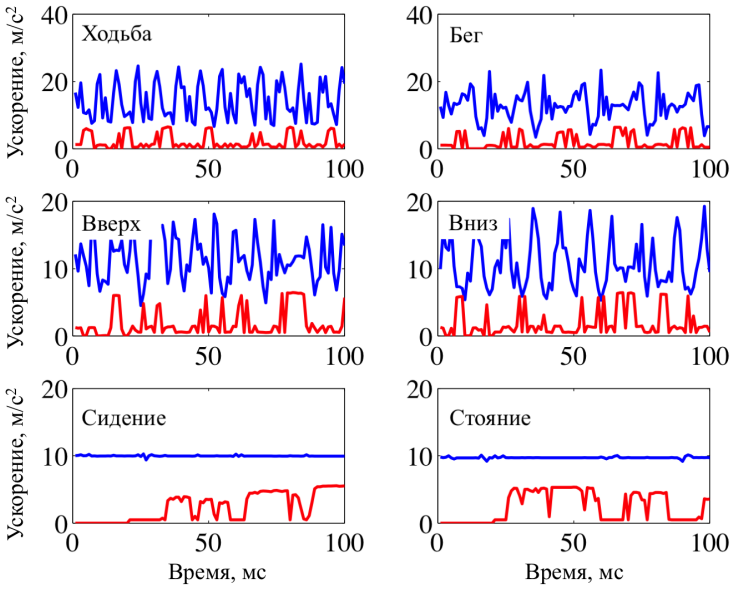


Рис. 4: Примеры временных рядов измерений акселерометра для разных видов физической активности

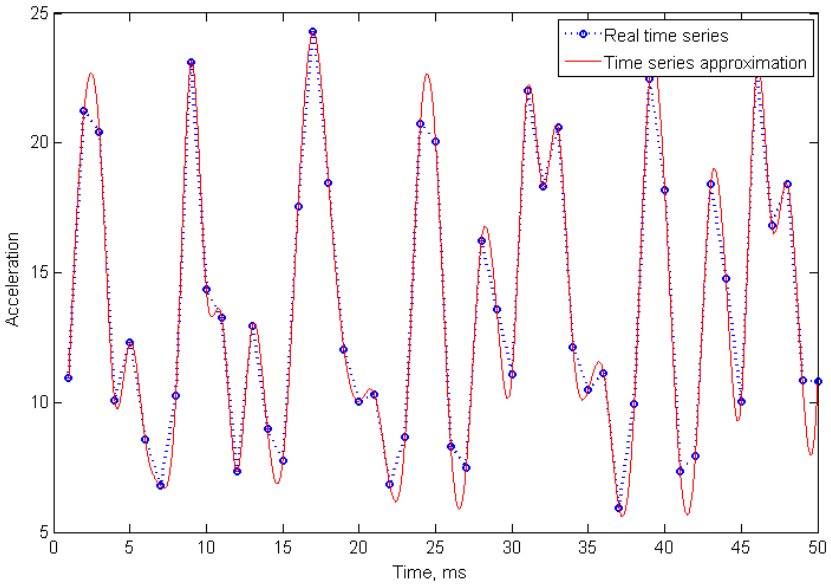


Рис. 5: Диаграмма рассеяния для признаков 6 и 2, по которой выдвигаются предположения о качестве классификации

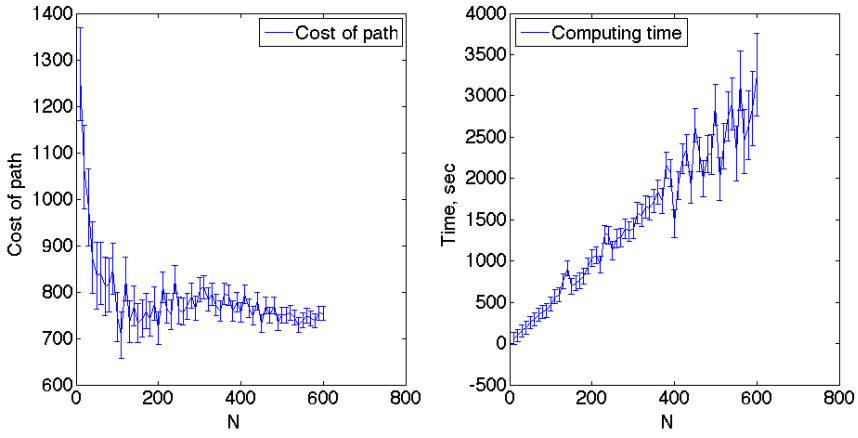


Рис. 6: Зависимость времени работы и значения расстояния от количества узлов сплайна.